# Predictive Modeling of Wildfires in the United States

Lang Qin
*College of Art & Science*
*University of Washingotn*
Seattle, U.S.
petertsing@outlook.com

Wenqian Shao
*College of Liberal Arts*
*University of Minnesota*
Minneapolis, U.S.
shao0089@umn.edu

Guofei Du
*School of Mathematics and Statistics*
*Southwest University*
Chongqing, China
2473471818@qq.com

Junlin Mou
*Art and Science Faculty*
*University of Toronto*
Toronto, Canada
junlin.mou@mail.utoronto.ca

Ran Bi
*College of Art & Science*
*University of Oregon*
Eugene, U.S.
br5265@hotmail.com

***Abstract*—This research utilizes wildfire records between 1911 and 2015 to train various models to predict fire size through using temperature, wind, humidity, and precipitation as features. Our results show 1) Decision Tree based Classifier outperforms both linear and ridge regression 2) Government entities can leverage our methodology to manage wildfires more efficiently, effectively, and =decreasing monetary damages.**

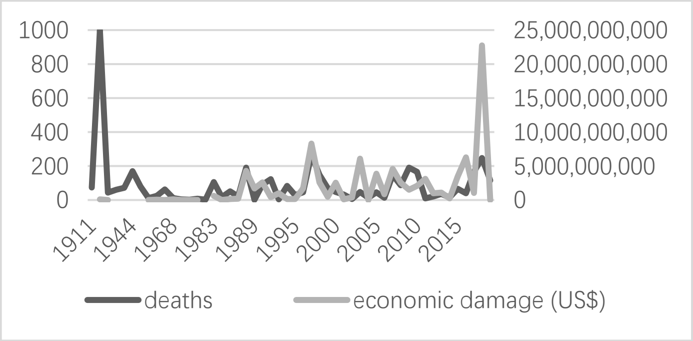
***Keywords*—wildfire, meteorological factors, predictive models, machine learning, learning regression, ridge regression, decisiontreeclassifier**

## I. Introduction

Wildfires, as one of the major natural disasters, have greatly impacted the normal life of the public. Because they occur in the wild, wildfires are often difficult to control. So every time a wildfire occurs, there will be a lot of damage and loss. It can affect the hydrological and geomorphic processes of river basins by changing elements like nitrogen in the water [1]. Huge impacts on the land such as soil erosion, accelerated deposition [2] and increased hydrophobicity [3]. According to the National Interagency Fire Center, between January 1 and September 8, 2020 there were 41,051 wildfires in the United States. This number in the same period in 2019 is 35,386 [4]. Fig.1 shows that the wildfires bring huge global economic damage: from 1918 to 2019, the economic damage has increased from 100 million dollar to over 20 billion, and every year, there are people killed by wildfires. Due to the serious impact, large numbers and high annual increases of wildfires indicate they have a great impact and need to be understood and prevented.

With the understanding of the significant impact on the society and nature brought by wildfires, it is worth mentioning that the function and trend of wildfires over the years. As a type of natural disaster, wildfires have typically caused great property damage and loss of life. This doesn't mean that wildfires are an entirely negative disaster; as part of the ecological chain, the soil is more fertile after a wildfire, and vegetation was more susceptible to insect infestation prior to a wildfire [5]. Notably, wildfires play an indispensable role in affecting the surface and atmosphere of the Earth for “over 350 million years” [6]. In fact, “fire is an important driver of change in the most forests, savannah, and prairie ecosystems and fire-altered organic matter convey numerous functions in soils of fire-maintained terrestrial ecosystems” [7]. Yet such essential and useful natural behaviour has been proven to accelerate global warming. Research shows that recent climate changes that create warmer, drier conditions, increased drought, and a more extended fire season are boosting increases in wildfire risk [8]. Particularly in the U.S., the project demonstrates that an average 1 Celsius degree increase would increase the median burned area per year as much as 600 percent in some types of forests, which results in cycling negative effect on global warming [9]. Hence, considering wildfire’s natural function and influence on global climate change, the purpose is not to eliminate wildfire, but to contain it.

Fig. 1. Global economic damage and the number of deaths from wildfires 1911-2019



Then it comes to a long-standing problem of how to prevent wildfires: from the old days when people controlled fires simply by dropping water or by cutting down quarantine belts, to the more recent times when people controlled fires artificially by building models [10] to predict the location, probability, extent, size, and direction [11] of fires [12] and to bring economic benefits. All these changes are due to the advancement of science and technology, which proves that the prevention and control of wildfire can only be improved by the innovation of science and technology.

People tried to predict wildfires through many different methods and focusing on different areas. One of the major techniques is to look at the spatial distribution to map wildfire patterns and therefore determine the probability of wildfires ignited [13]. Besides predicting wildfires based on geographical locations, another popular method is to look at drivers for wildfires such as precipitation, gas composition, and human activities to build prediction models [14]. There are also studies that combine both methods [15]. There is a dynamic model WIFIRE that uses monitoring sensors, satellite, even twitter to give real-time, data-driven predictions [16].

Additionally, many previous studies demonstrated that it is feasible to predict the scale of a forest wildfire at the beginning of its occurrence using meteorological information. Taking the meteorological factors as input values, long short-term memory (LSTM) which is one of the classification methods is implemented to establish prediction models, and exhibited the highest accuracy [17]. Another research uses the road density as a surrogate for human access and behavior, and the result shows that areas which are close to denser roads are more likely to get human-caused wildfires [18]. To predict events of large acreages burned by wildfires, logistic regression can be applied, and the goodness of predictions is measured by specificity, sensitivity, and correctness using a cross-validation method [19]. Logistic regression can be used in combination with land cover, vegetation index, topographic and socioeconomic information to characterize the spatial pattern of the fire occurrence [20].

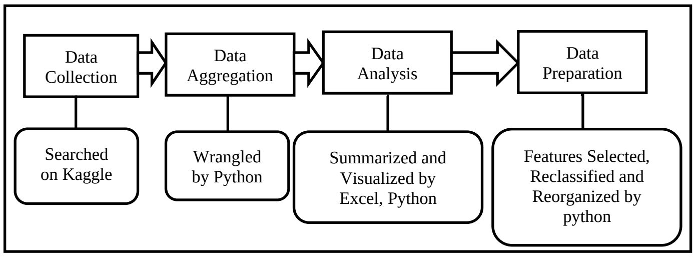
Based on previous scholars' achievements in modeling and prediction, with more comprehensive data sets provided, this study will analysis and further utilize machine learning to construct predictive models with respect to the following elements: temperature, wind, humidity, and precipitation before the occurrence of the wildfire. With the model, the government and relevant organizations will receive a better view of possible development of wildfire and thus better constrain wildfires damage to the climate and society on a balanced point.

## II. Data

### A. Data Processing Steps

We processed the data according to the following steps, data collection, data aggregation, data analysis, and data preparation. (Figure 2)

Figure. 2. Data Process Flow Chart


Data Collection
Data Aggregation
Data Analysis
Data Preparation
Searched on Kaggle
Wrangled by Python
Summarized and Visualized by Excel, Python
Features Selected, Reclassified and Reorganized by python

### B. Data Source

The dataset used for this research is found on Kaggle, U.S. Wildfire Data, 2020, which is a subset of 1.8 Million U.S wildfires [21] joined with other related databases historical weather data at a specific latitude and longitude [22], historical vegetation data [23]. A metric is representing the measure of the remoteness of a fire using city latitude and longitude database [24].

### C. Data Overview

The datasets contain 55367 observations of wildfires in the United States between 1991 and 2015 which includes 43 different variables. However, rather than including all the variables, this research will only focus on the variables presented on the following table since other variables are not relevant for the research topic.

Table 1. Variables Description

| Name of Variables | Description |
|---|---|
| fire_size | Size of fire in acres |
| fire_size_class | Class of fire size (A-G) |
| Temp_pre_30 | Temperature in °C at the location of fire up to 30 days prior |
| Temp_pre_15 | Temperature in °C at the location of fire up to 15 days prior |
| Temp_pre_7 | Temperature in °C at the location of fire up to 7 days prior |
| Temp_cont | Temperature in °C at the location of fire up to day the fire was contained |
| Wind_pre_30 | Wind speed in m/s at the location of fire up to 30 days prior |
| Wind_pre_15 | Wind speed in m/s at the location of fire upto 15 days prior |
| Wind _pre_7 | Wind speed in m/s at the location of fire up to 7 days prior |
| Wind _cont | Wind speed in m/s at the location of fire up to day the fire was contained |
| Hum_pre_30 | Humidity in % at the location of fire up to 30 days prior |
| Hum _pre_15 | Humidity in % at the location of fire upto 15 days prior |
| Hum _pre_7 | Humidity in % at the location of fire up to 7 days prior |
| Hum _cont | Humidity in % at the location of fire up to day the fire was contained |
| Prec _pre_30 | Precipitation in mm at the location of fire up to 30 days prior |
| Prec _pre_15 | Precipitation in mm at the location of fire upto 15 days prior |
| Prec _pre_7 | Precipitation in mm at the location of fire up to 7 days prior |
| Prec _cont | Precipitation in mm at the location of fire up to day the fire was contained |

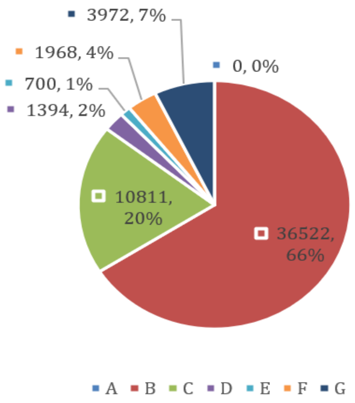
The fire size's classification is determined by the conflagration area in acres.

Table 2. Fire Size Classification

| Fire Size Class | Conflagration Area (Acres) |
|---|---|
| A | <=0.25 |
| B | 0.26-9.9 |
| C | 10.0-99.9 |
| D | 100-299 |
| E | 300-999 |
| F | 1000-4999 |
| G | >=5000 |

Fig. 3 shows the majority of wildfires' are class B (66%), one fifth in C, and no fire of size A is recorded in this dataset.

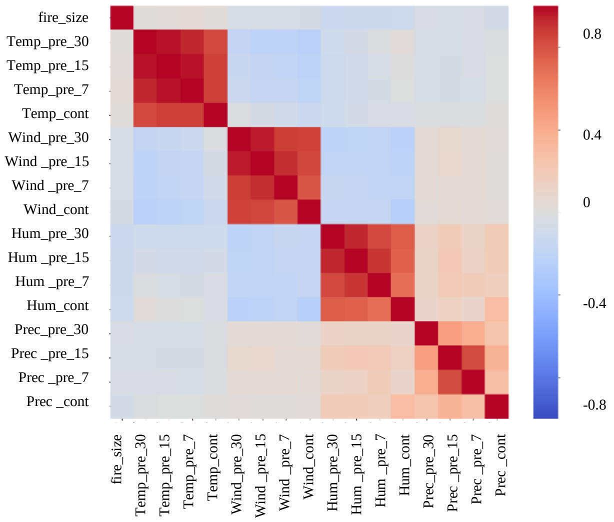
Figure 3. Fire Record Counts by Fire Size Class



To be noted, data-wrangling resulted in 7786 valid cases of wildfire with complete information used for this analysis.

The correlation matrix (Figure 4) shows correlation between the variables used in this study. strong positive correlations (>0.8) can be observed for each category (temperature, humidity, wind speed and precipitation). All other correlations have an absolute value below 0.4. Fire size has a positive correlation with temperature and a negative correlation with precipitation and humidity.

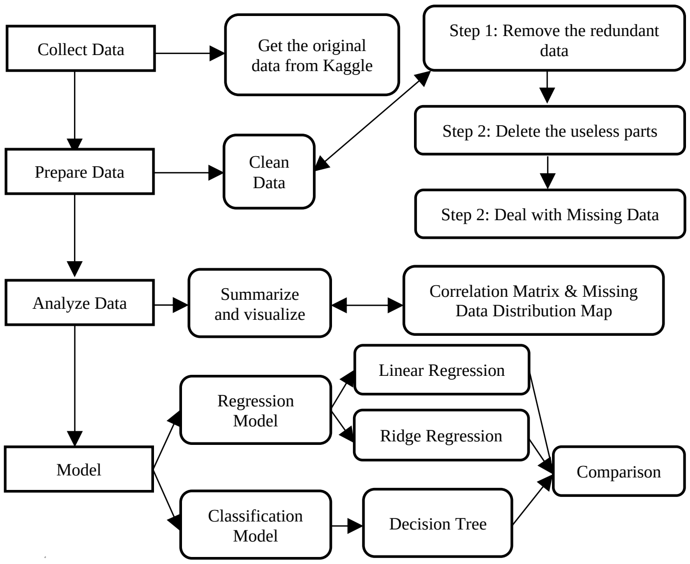
Figure 4. Correlation Map



## III. Methodology

This research consists of following steps: data preparation, data analysis and modeling (Figure 5). To select the most suitable predictive model, this research constructed two regression models: Linear Regression and Ridge Regression, one classification model: Decision Tree Classifier, and we divided the data by 70% for training and 30% for test to evaluate the models.

Figure 5. Methodology


Collect Data
Get the original data from Kaggle
Step 1: Remove the redundant data
Prepare Data
Clean Data
Step 2: Delete the useless parts
Step 2: Deal with Missing Data
Analyze Data
Summarize and visualize
Correlation Matrix & Missing Data Distribution Map
Linear Regression
Regression Model
Ridge Regression
Model
Comparison
Classification Model
Decision Tree

Note: This diagram shows the flow of the study. We get the data from Kaggle, and then clean the data to prepare for the analysis. Construct three models respectively, and compare them.

### A. *Linear Regression*

Through linear regression, we can obtain the quantitative relationship of interdependence between two or more variables. A linear regression analysis in which the least squares approximation is used to fit the relationship between one or more independent and dependent variables for modeling. In general, there will be multiple factors affecting the dependent variable $y$. Assuming that there are $x_1, x_2, x_3, \ldots, x_k$ , k factors, the following linear relationship can be considered:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon$$

And then do n times independent observations to obtain n sets of observed values $(x_{t1}, x_{t2}, \ldots, x_{tk})$ and they satisfy:

$$y = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \ldots + \beta_{tk} x_k + \epsilon_t$$

Hence we get $Y = X\beta + \epsilon$ , solve the $\beta$ using $\hat{\beta} = (X^{\top}X)^{-1}X^{\top}Y$.

In our study, we want to find a correlation between the size of wildfires and the following elements: temperature, wind, humidity, and precipitation before the occurrence of wildfires. There are four factors that may affect the dependent variable. Through the above model, we can predict and estimate the size of the wildfires.

### B. *Ridge Regression*

Ridge regression is proposed to solve the instability problem of the numerical solutions of linear regression. The normal linear regression equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon$$

$$\omega = (\, \beta_0 \quad \beta_1 \quad \beta_2 \quad \beta_3 \quad \cdots \quad \beta_k \,)$$

Its loss function:

$$\min_{\omega} \|X\omega - y\|_2^2$$

When we use the above function to calculate the parameters of the model, the absolute value of the parameters tends to become very large.

To limit the value of the parameter $\omega$, we can impose a penalty term on the size of the coefficients. This process is called regularization, and the penalized residual sum of the squares is:

$$\underset{\omega}{min}\|X\omega - y\|_2^2 + \alpha\|\omega\|_2^2$$

The extra term is the parameters' sum of squares. The complexity parameter $\alpha \geq 0$ controls the amount of shrinkage: the larger the value of $\alpha$, the greater the amount of shrinkage and thus the coefficients become more robust to collinearity. In our study, we use $\alpha = 0.5$ to balance the variance and bias of the model, and we compare the result with linear regression to increase the precision of the prediction.

### C. *DecisionTreeClassifier*

Decision tree is that the computer determines the possibility of the results in various situations according to specific filtering logic. In our project, we hope that we can determine the size of the wildfire and the possibility of its occurrence through various classification factors. As a feature of decision tree, we can select many attributes as our decision points. For example, humidity, wind speed, temperature in 30 days. At the same time, in our data, there are classes of fire-size (A-G) as the result of these factors, and this is our category. Our code gives the possibility of various fire-size to happen through known humidity, wind speed and temperature.

### D. *Comparison*

To compare the three models, we will score them based on the test results, and we use different methods to score models of regression and classification.

For regression, we use $R^2$ score to compute the coefficient of determination. $R^2$ is defined as $(1 - \frac{u}{v})$, where the $u$ means the residual sum of squares, and $v$ is the total sum of squares. The best possible score is 1, and in our research linear regression's score is approximately 0.03 and ridge regression's score is about 0.03, which means models based on regression do not perform well in our dataset.

For classification, we introduce accuracy score to evaluate the precision. In the test, if one sample's predicted value $\hat{y}_i$ equals to the corresponding true value $y_i$, the indicator function equals to 1, $1(x) = 1$. Thus, the fraction of correct predictions over $n$ samples is:

$$accuracy(\hat{y}_i, y_i) = \frac{1}{n}\sum_{i=0}^{n-1} 1\,(\hat{y}_i = y_i)$$

The best possible accuracy is 1.0, and our model's accuracy score is approximately 0.7 which performs much better than the regression models.

## IV. Result

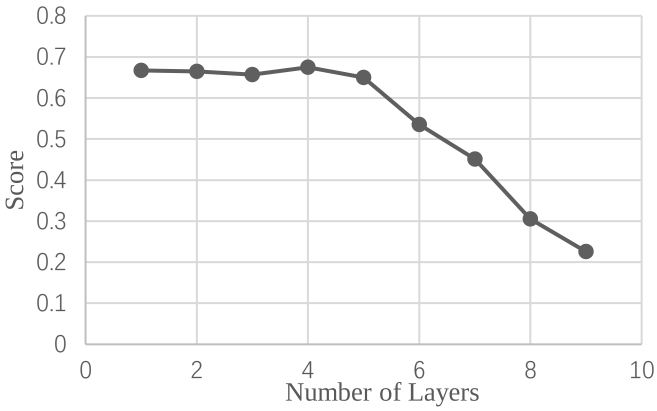
We compared each predictive model according their prediction score. Regression models, including Linear Regression and Ridge Regression, did not yield good results. However, the classification model, Decision Tree, met our target score of nearly 0.70 (Table 3). To optimize the classification model, we tested the decision layers from 1 to 9 and concluded that, at max learning depth equals 4, the classification model would show highest score (Figure 6).

Table 3. Scores of Models

| **Model Type** | **Score (1 for Best)** |
|---|---|
| Linear Regression | 0.03845 |
| Ridge Regression | 0.03242 |
| Decision Tree | 0.67467 |

Note: This table lists the score of each machine learning model in prediction of the size of wildfires.

Figure 6. Score vs. Max Depth



Note: This figure shows the score of classification tree at different number of layers.

Since the decision tree algorithm exceeded both regression models, we chose the classification model as our final predictive model on the size of wildfire. Noticeably, the correlation coefficients as the score of linear regression model and ridge regression model are below 0.05, which suggests that both models are not suitable for constructing numerical predictive model for the size of wildfires.

## V. Conclusion

A predictive model to predict the size of wildfires is beneficial for constraining wildfires. A decision tree is an effective predictive model on the class of the wildfire size, based on previous information of temperature, humidity, wind, and precipitation. The outbreak of wildfires continues to grow each year, and social damage shows increasing trends. In recent years, financial damage relevant to wildfires has increased from $5 billion to  $20 billion between the years of 2005 and 2015. Because there exists limited human resources and equipment to contain each wildfire, it would prudent to deploy more resources to fires whose class prediction is higher to optimize resources and limit damage. The idea of any predictive model is to firstly prevent and then constrain the damage of wildfire. Considering the fact that wildfire is an indispensable part for nature, it is important to find the balanced point between the necessity of the nature and damage to human society. With predictive models in this research, classification model by Decision Tree, the specialists will receive a broader and earlier view of the wildfire, and then, find the balance between the human and nature.

For future research we plan on addressing some of these problems. Regression predictive models are constructed linearly. A more advanced model may be better at observing and find the possible correlations between many variables. With a comprehensive understanding of the wildfires, we can improve the model by using weighted values based on the influence of each factor. Additionally, geographical and biological factor will be taken into consideration to construct a more realistic and comprehensive model.

## References

[1] Mahat, Vinod, Anderson, Axel, & Silins, Uldis. (2015). Modelling of wildfire impacts on catchment hydrology applied to two case studies. Hydrological Processes, 29(17), 3687-3698.

[2] Wondzell, S., & King, J. (2003). Postfire erosional processes in the Pacific Northwest and Rocky Mountain regions. Forest Ecology and Management., 178(1–2), 75-87.

[3] Scott, D., & Van Wyk, D. (1990). The effects of wildfire on soil wettability and hydrological behaviour of an afforested catchment. Journal Hydrology, 121(1–4), 239-256.

[4] National Interagency Fire Center. (2020, September 24). Report. https://www.nifc.gov/fireInfo/nfn.htm3.

[5] Andrews, Patricia, et al. (2017). "Predicting Wildfires." SCIENTIFIC AMERICAN. https://www.fs.fed.us/rm/pubs other/rmrs 2007 andrews p001.pdf.

[6] Doerr, S.H. & Sant´ın, Cristina. (2016). Global trends in wildfire and its impacts: Perceptions versus realities in a changing world. Philosophical Transactions of the Royal Society B: Biological Sciences. 371.20150345. 10.1098/rstb.2015.0345.

[7] Pingree, M. R., & DeLuca, T. H. (2017, August 08). Function of Wildfire-Deposited Pyrogenic Carbon in Terrestrial Ecosystems. Retrieved from Https://doi.org/10.3389/feK. Elissa, "Title of paper if known," unpublished.

[8] USGCRP, (2017): Climate Science Special Report: Fourth National Climate Assessment, Volume I [Wuebbles, D.J., D.W. Fahey, K.A. Hibbard, D.J. Dokken, B.C. Stewart, and T.K. Maycock (eds.)]. U.S. Global Change Research Program, Washington, DC, USA, 470 pp, doi: 10.7930/J0J964J6.M.

[9] James M. Vose, David L. Peterson, & Toral PatelWeynand. (2012). Effects of Climatic Variability and Change on Forest Ecosystems: A Comprehensive Science Synthesis for the U.S. Forest Sector [Scholarly project]. PNWGTR870, General Technical Report. United States Department of Agriculture, Forest Service, Pacific Northwest Research Station. In Https://www.fs.fed.us/pnw/pubs/pnw gtr870/pnw gtr870.pdf

[10] .Carl C. Wilson & James B. Davis. Forest Fire Laboratory at Riverside and Fire Research in California: Past, Present, and Future.United States Department of Agriculture. https://www.fs.fed.us/psw/publications/documents/pswgtr105/pswgtr105.pdf

[11] National Science and Analysis Team. "Scientific Basis for Modeling Wildland Fire Management: The Phase II Report of the National Science and Analysis Team." (17 Jan. 2012), www.forestsandrangelands.gov/documents/strategy/reports/phase2/NSAT Phase 2 Summary

[12] Gorte, Ross W., and Kelsi Bracmort. (7 Mar. 2012). "ForestFire/Wildfire Protection." Congressional Research Service, fas.org/sgp/crs/misc/RL30755.pdf.

[13] Hardtke, L. A., Valle, H. F., & Sione, W. (2011). Spatial distribution of wildfire risk in the Monte biome (Patagonia, Argentina). Journal of Maps, 7(1), 588-599. doi:10.4113/jom.2011.1184

[14] Pilliod, D. S., Welty, J. L., & Arkle, R. S. (2017). Refining the cheatgrass-fire cycle in the Great Basin: Precipitation timing and fine fuel composition predict wildfire trends. Ecology and Evolution, 7(19), 8126-8151. doi:10.1002/ece3.3414

[15] Guo, F., Su, Z., Wang, G., Sun, L., Lin, F., & Liu, A. (2016). Wildfire ignition in the forests of southeast China: Identifying drivers and spatial distribution to predict wildfire likelihood. Applied Geography, 66, 12-21. doi:10.1016/j.apgeog.2015.11.014

[16] Crawl, D., Block, J., Lin, K., & Altintas, I. (2017). Firemap: A Dynamic Data-Driven Predictive Wildfire Modeling and Visualization Environment. Procedia Computer Science, 108, 2230-2239. doi:10.1016/j.procs.2017.05.174

[17] Liang Hao, Meng Zhang, & Hailan Wang. (2019). A Neural Network Model for Wildfire Scale Prediction Using Meteorological Factors. IEEE Access 7: 176746–55.

[18] Marchal Jean, Steve G. Cumming, & Eliot J. B. McIntire. (2017). Exploiting Poisson Additivity to Predict Fire Frequency from Maps of Fire Weather and Land Cover in Boreal Forests of Qu´ebec, Canada. Ecography 40(1): 200–209.

[19] Chu Pao-Shin, Weiping Yan, & Francis Fujioka. (2002). Fire-Climate Relationships and Long-Lead Seasonal Wildfire Prediction for Hawaii. International Journal of Wildland Fire 11(1): 25.

[20] Zhang Yang, Samsung Lim, & Jason John Sharples. (2016). Modelling spatial patterns of wildfire occurrence in South-Eastern Australia. Geomatics, Natural Hazards and Risk 7(6): 1800–1815.

[21] Short, Karen C. 2017. Spatial wildfire occurrence data for the United States, 1992-2015 [FPAFOD20170508]. 4th Edition. Fort Collins, CO: Forest Service Research Data Archive. https://doi.org/10.2737/RDS-2013-0009.4

[22] NOAA National Centers for Environmental Information (2001): Integrated Surface Hourly [1992-2015] - ftp://ftp.ncdc.noaa.gov/pub/data/noaa/

[23] Meiyappan, Prasanth, and Atul K. Jain. "Three distinct global estimates of historical land-cover change and land-use conversions for over 200 years." Frontiers of Earth Science 6.2 (2012): 122-139.

[24] Meiyappan, Prasanth, and Atul K. Jain. "Three distinct global estimates of historical land-cover change and land-use conversions for over 200 years." Frontiers of Earth Science 6.2 (2012): 122-139.